PORTABLE PHYSICIAN

A mini-project report submitted to Rashtrasant Tukadoji Maharaj Nagpur University in partial fulfilment for the award of degree of

Bachelor of Engineering

In

Computer Science and Engineering

By

ANIRUDHA AGRAWAL

HARSH GOYAL

RISHAB SHUKLA

SURBHI KHATRI

SHWETA CHAUDHARY

Under the supervision of

Dr. R. S. HABLANI

Department of Computer Science and Engineering,
Shri Ramdeobaba College Of Engineering And Management,
Nagpur-440013

(An Autonomous Institution of Rashtrasant Tukadoji Maharaj Nagpur University)

November 2019

SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT, NAGPUR

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the Thesis on

"PORTABLE PHYSICIAN"

is a bonafide work of

ANIRUDHA AGRAWAL

HARSH GOYAL

RISHAB SHUKLA

SURBHI KHATRI

SHWETA CHAUDHARY

submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillment of the award of a Degree of Bachelor of Engineering, in Computer Science and Engineering. It has been carried out at the Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur during the academic year 2019-20.

Date : 2 November, 2019
Place: Nagpur

Dr. R. S. HABLANI
Project guide
Department of Computer Science
and Engineering

Dr. M.B. Chandak
H.O.D
Department of Computer Science
and Engineering

Dr. R.S. Pande

Principal

DECLARATION

We, hereby declare that the project titled **"PORTABLE PHYSICIAN"** submitted herein, has been carried out in the Department of Computer Science and Engineering of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree /diploma at this or any other institution / University.

Date : 2 November, 2019

Place: Nagpur

ANIRUDHA AGARWAL

Roll No. 46

HARSH GOYAL

Roll No. 54

SURBHI KHATRI

Roll No. 18

SHWETA CHAUDHARY

Roll No. 15

RISHABH SHUKLA

Roll No. 92

Approval Sheet

This report entitled

"PORTABLE PHYSICIAN"

by

ANIRUDHA AGRAWAL

HARSH GOYAL

RISHAB SHUKLA

SURBHI KHATRI

SHWETA CHAUDHARY

is approved for the degree of Bachelor in Computer Science and Engineering.

Name & signature of Supervisor	Name & signature of External. Examiner
----------------------------	----------------------------
----------------------------	----------------------------
Name & signature RRC Members	Name & signature of HOD
----------------------------	----------------------------
----------------------------	----------------------------

Date: 2 November 2019

Place: Nagpur

ACKNOWLEDGEMENT

Every project small or big is not possible without the guidance and help of various numbers of people in various fields who have always given their valuable advice and support. We sincerely appreciate the inspiration, guidance, dedication and positivity shown by all those people who in some or the other way have contributed in making this project a success.

We are grateful to our Project Guide Prof. R.S Hablani who was always there to solve our every small doubt about the project. He not only gave us the right path towards approaching our goal, but was also there to provide moral support and valuable inputs.

We forward our thanks to Prof. A.R. Raipurkar, Project Co-ordinator for providing support and guidance for the project completion.

We are indebted to our Head of Department, Dr. M. B. Chandak who gave us a positive response and provided us various approaches through which the project could be implemented and developed further.

We are thankful to all our panelists and supervisors who were present during the project presentation and submission for their valuable comments and advice which also helped us in improving our presentation and communication skills.

By –

ANIRUDHA AGRAWAL 46

HARSH GOYAL 54

RISHAB SHUKLA 92

SURBHI KHATRI 18

SHWETA CHAUDHARY 15

ABSTRACT

Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from biomedical literature has gained popularity among researchers, and deep learning has boosted the development of effective biomedical text mining models. However, directly applying the advancements in NLP to biomedical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora. Here we investigate how the recently introduced pre-trained language model BERT can be adapted for biomedical corpora.

Results: We developed our project which uses Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. Our analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts. Which can be used to give answers to simple queries. We made an application which can take input in native language like Hindi, Marathi, Tamil, Telugu etc from the user of the application and and respond back to user according to his/her symptoms in the same language.

LIST OF FIGURES

INDEX

CHAPTER 1

INTRODUCTION

1.1 Background

The volume of biomedical literature continues to rapidly increase. On average, more than 3000 new articles are published every day in peer-reviewed journals, excluding pre-prints and technical reports such as clinical trial reports in various archives. PubMed alone has a total of 29M articles as of January 2019. Reports containing valuable information about new discoveries and new insights are continuously added to the already overwhelming amount of literature. Consequently, there is increasingly more demand for accurate biomedical text mining tools for extracting information from the literature.

1.2 Problem Definition

Create AI based doctor that can diagnose everyday acute diseases like common cold ,flu etc,based on simple questions.

1.3 Need of the system

We hope that our Project can lead to significant advances in medical technologies which can diagnose at the level of experts, towards improving healthcare access in parts of the world where access to skilled doctors is limited.

1.4 Objectives of the system

The objectives are as follows:

1.To take symptoms of disease as input in native language.

2.To take relevant information required for accurate diagnosis.

3.To predict accurate disease the user might be suffering from.

4.To suggest the most cost efficient way of treatment.

CHAPTER 2
REVIEW OF LITERATURE

2.1 Definition of the system

Supply of doctors is limited in India specially in smaller towns and villages making provision of healthcare is difficult to a large no of people. Telemedicine and other solutions in the past have also struggled to scale up due to this problem. Now in the age of digital assistants like google and Alexa we can create an AI based doctor that can diagnose everyday acute diseases like common cold flu etc. based on some simple questions. To make our model more reachable we are giving support of regional languages using google api.

2.2 Related Works

Recent progress of biomedical text mining models was made possible by the advancements of deep learning techniques used in natural language processing (NLP). For instance, Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) have greatly improved performance in biomedical named entity recognition (NER) over the last few years (Giorgi and Bader, 2018; Habibi et al., 2017; Wang et al., 2018; Yoon et al., 2019). Other deep learning based models have made improvements in biomedical text mining tasks such as relation extraction (RE) and question answering (QA) (Wiese et al., 2017).

2.3 Features of the System

This is the first domain-specific BERT based model pretrained on biomedical corpora for days on NVIDIA GPUs. We show that pre-training BERT on biomedical corpora largely improves its performance. It obtains higher F1 scores in biomedical NER and biomedical RE, and a higher MRR score in biomedical QA than the current state-of-the-art models. Compared with most previous biomedical text mining models that are mainly focused on a single task such as NER or QA, our model achieves state-of-the-art performance on various biomedical text mining tasks, while requiring only minimal architectural modifications.

2.4 Limitations of existing system

Directly applying state-of-the-art NLP methodologies to biomedical text mining has limitations. First, as recent word representation models such as Word2Vec, Elmo and BERT are trained and tested mainly on datasets containing general domain texts (e.g. Wikipedia), it is difficult to estimate their performance on datasets containing biomedical texts. Also, the word distributions of general and biomedical corpora are quite different, which can often be a problem for biomedical text mining models. As a result, recent models in biomedical text mining rely largely on adapted versions of word representations.

2.5 Advantages of current system

In this system, we hypothesize that current state-of-the-art word representation models such as BERT need to be trained on biomedical corpora to be effective in biomedical text mining tasks. Previously, Word2Vec, which is one of the most widely known context independent word representation models, was trained on biomedical corpora which contain terms and expressions that are usually not included in a general domain corpus (Pyysalo et al., 2013). While ELMo and BERT have proven the effectiveness of contextualized word representations, they cannot obtain high performance on biomedical corpora because they are pre-trained on only general domain corpora. As BERT achieves very strong results on various NLP tasks while using almost the same structure across the tasks, adapting BERT for the biomedical domain could potentially benefit numerous biomedical NLP researches. Other benefits are as follows :

1. Ease of use for rural population.

2.Free access to medical health-care.

3.24x7 access to health-care.

4.Availability of health-care also where physical doctors are not available.

CHAPTER 3

ANALYSIS AND DESIGN

In this chapter, the procedure for implementation of the project and a brief explanation of why it will be useful for implementing the proposed system is included, and a brief description of the current system development approach is counted.

3.1 Methodology

The project development methodology is as follows :

- First we will take the symptoms or query as input in the native language of the user.
- Then we will translate the native language into English language to pass to the model.
- We are going to web scrape the data to get symptoms and the corresponding disease and medication.
- From the data collected from various web sources we are going to train our LSTM model.
- Then we will report the output of this model to the user.
- It will be able to recommend medications to the user.
- All this we will integrate on our android app for portablity of the user.

3.2 Flow of Control

User will be facilitate with login functionality.

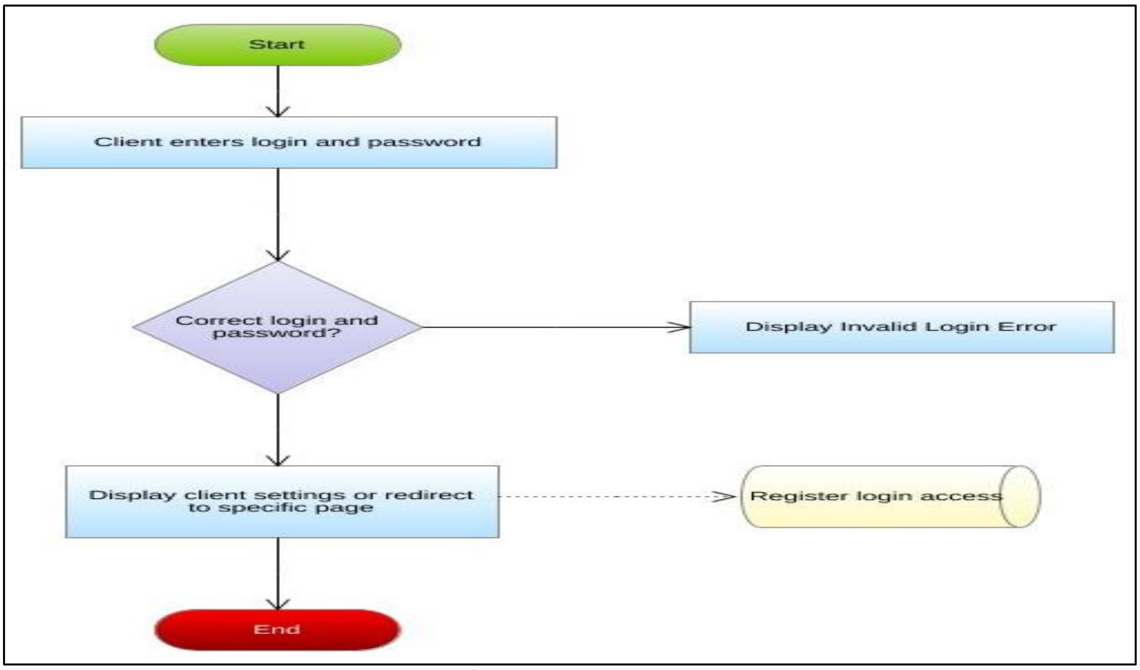


fig 3.1 Login Flow

Our baseline will use a Bidirectional Encoder Representations from Transformers (BERT) to detect the cause of problem .

- First the user will speak the cause in its native language
- Then the voice will be first converted to text using the speech to text api.
- Then the text will be converted to English text which is then ready to be inserted in the model .
- Model returns the answers for the query or subsequent questions which then as a whole returns the final solution to the query.
- Then this solution which is in English is converted to the regional language .
- And if required then this text is converted to speech using google api.

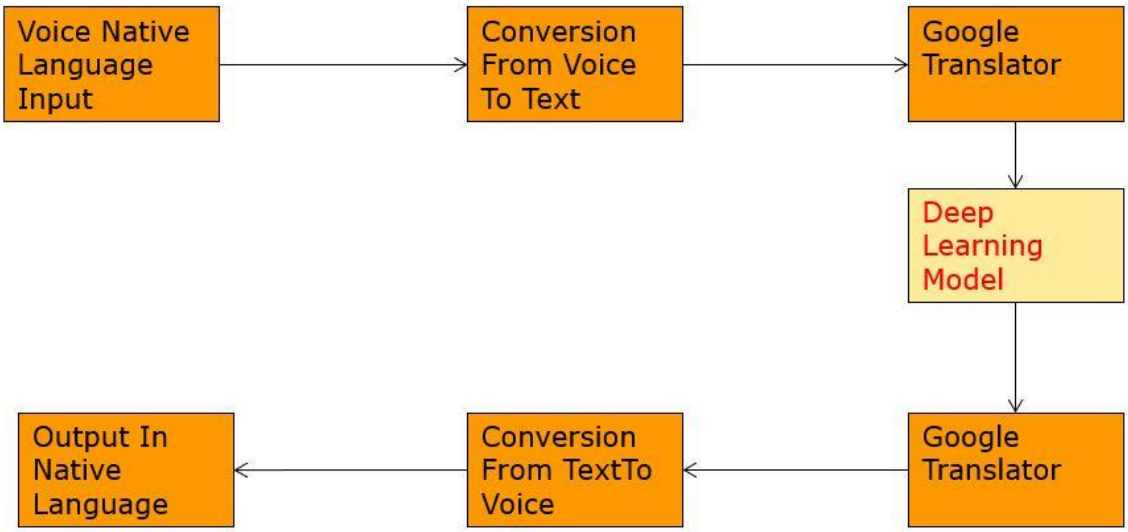


Fig 3.2 Activity flow diagram of System

3.3 Process

This basically has the same structure as BERT Bidirectional encoder representations from transformers Learning word representations from a large amount of unannotated text is a long-established method. While previous models e.g. Word2Vec, GloVe focused on learning context independent word representations, recent works have focused on learning context dependent word representations. For instance, ELMo (Peters et al., 2018) uses a bidirectional language model, while CoVe (McCann et al., 2017) uses machine translation to embed context information into word representations. BERT (Devlin et al., 2019) is a contextualized word representation model that is based on a masked language model and pretrained using bidirectional transformers (Vaswani et al., 2017). Due to the nature of language modelling where future words cannot be seen, previous language models were limited to a combination of two unidirectional language models (i.e. left-to-right and right-to left). BERT uses a masked language model that predicts randomly masked words in a sequence, and hence can be used for learning bidirectional representations. Also, it obtains state-of-the-art performance on most NLP tasks, while requiring minimal task-specific architectural modification. According to the authors of BERT, incorporating information from bidirectional representations, rather than unidirectional representations, is crucial for representing words in natural language. We hypothesize that such bidirectional representations are also critical in biomedical text mining as complex relationships between biomedical terms often exist in a biomedical corpus . Due to the space limitations, we refer readers to Devlin et al. (2019) for a more detailed description of BERT.

Pre-training Our Model

As a general purpose language representation model, BERT was pretrained on English Wikipedia and Books Corpus. However, biomedical domain texts contain a considerable number of domain-specific proper nouns and terms (e.g. transcriptional, antimicrobial), which are understood mostly by biomedical researchers. As a result, NLP models designed for general purpose language understanding often obtains poor performance in biomedical text mining tasks. In this work, we pre-train model on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). The text corpora used for pre-training of this and the tested combinations of text corpora are listed in table . For computational efficiency, whenever the Wiki þ Books corpora were used for pre-training, we initialized it with the pre-trained BERT model provided by Devlin et al. We define it as a language representation model whose pre-training corpora includes biomedical corpora. For tokenization, it uses Word Piece tokenization, which mitigates the out-of-vocabulary issue. With Word Piece tokenization, any new words can be represented by frequent sub words .We found that using cased vocabulary (not lowercasing) results in slightly better performances in downstream tasks. Although we could have constructed new Word Piece vocabulary based on biomedical corpora, we used the original vocabulary of BERTBASE for the following reasons: (i) compatibility of this with BERT, which allows BERT pre-trained on general domain corpora to be re-used, and makes it easier to interchangeably use existing models based on BERT (ii) any new words may still be represented and fine-tuned for the biomedical domain using the original Word Piece vocabulary of BERT.

Fine-tuning

With minimal architectural modification, it can be applied to various downstream text mining tasks. We fine-tune pretrained model on the following three representative biomedical text mining tasks: NER, RE and QA. Named entity recognition is one of the most fundamental biomedical text mining tasks, which involves recognizing numerous domain-specific proper nouns in a biomedical corpus. While most previous works were built upon different combinations of LSTMs and CRFs, BERT has a simple architecture based on bidirectional transformers. BERT uses a single output layer based on the representations from its last layer to compute only token level BIO2 probabilities. Note that while previous works in biomedical NER often used word embeddings trained on PubMed or PMC corpora, It directly learns Word Piece embeddings during pre-training and fine-tuning. For the evaluation metrics of NER, we used entity level precision, recall and F1 score. Relation extraction is a task of classifying relations of named entities in a biomedical corpus. We utilized the sentence classifier of the original version of BERT, which uses a token for the classification of relations. Sentence classification is performed using a single output layer based on a token representation from BERT. We anonymized target named entities in a sentence using pre-defined tags such as @GENE$ or @DISEASE$. For instance, a sentence with two target entities (gene and disease in this case) is represented as "Serine at position 986 of @GENE$ may be an independent genetic predictor of angiographic @DISEASE$." The precision, recall and F1 scores on the RE task are reported. Question answering is a task of answering questions posed in natural language given related passages. To fine-tuned model for QA, we used the same BERT architecture used for SQuAD (Rajpurkar et al., 2016). We used the BioASQ factoid datasets because their format is similar to that of SQuAD. Token level probabilities for the start/end location of answer phrases are computed using a single output layer. However, we observed that about 30% of the BioASQ factoid questions were unanswerable in an extractive QA setting as the exact answers did not appear in the given passages. Like Wiese et al. , we excluded the samples with unanswerable questions from the training sets. Also, we used the same pre-training process of Wiese et al. (2017), which uses SQuAD, and it largely improved the performance of both BERT and our model. We used the following evaluation metrics from BioASQ: strict accuracy, lenient accuracy and mean reciprocal rank.

Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

Model	Corpus combination
BERT (Devlin *et al.*, 2019)	Wiki + Books
BioBERT (+PubMed)	Wiki + Books + PubMed
BioBERT (+PMC)	Wiki + Books + PMC
BioBERT (+PubMed + PMC)	Wiki + Books + PubMed + PMC

Structure of System Model

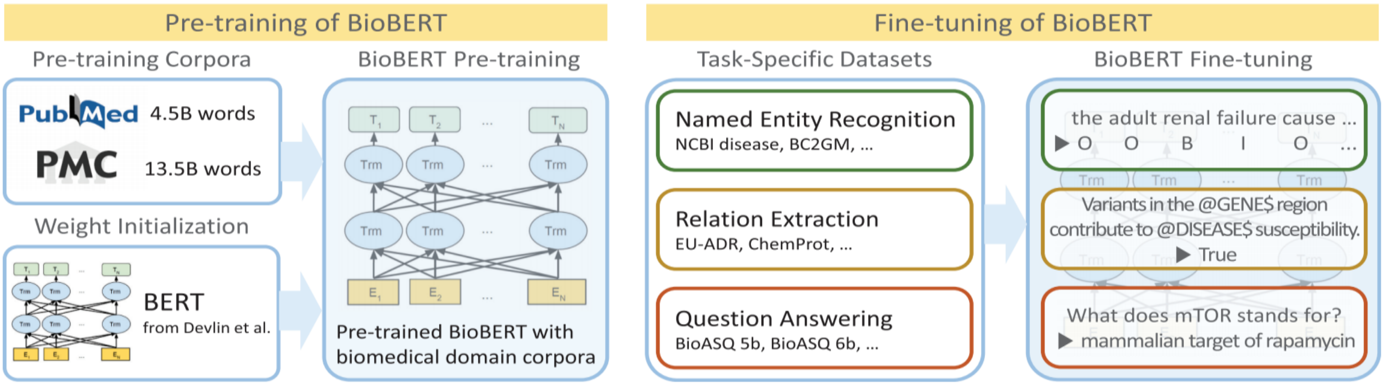


Fig 3.3 System Model Structure

CHAPTER 4
SYSTEM DESCRIPTION

1.1 Software tools

1. **Sublime Text**

   It is a sophisticated text editor for code, mark up and prose. You'll love the slick user interface, extraordinary features and amazing performance. Brackets is a text editor tool that makes it easy to design in the browser, it is crafted from the ground up for web designers and front-end developers

2. **Python 3.7**

   Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

3. **Word2Vec**

   Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

4. **Pandas**

   Pandas (software) In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

5. **NumPy**

   NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

6. **Beautiful Soup**

   Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

7. **HuggingFace Transformers**

   Transformers (formerly known as pytorch-transformers and pytorch-pretrained-bert) provides state-of-the-art general-purpose architectures (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32+ pretrained models in 100+ languages and deep interoperability between TensorFlow 2.0 and PyTorch.

8. **Tensorflow**

   TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

9. **Flutter**

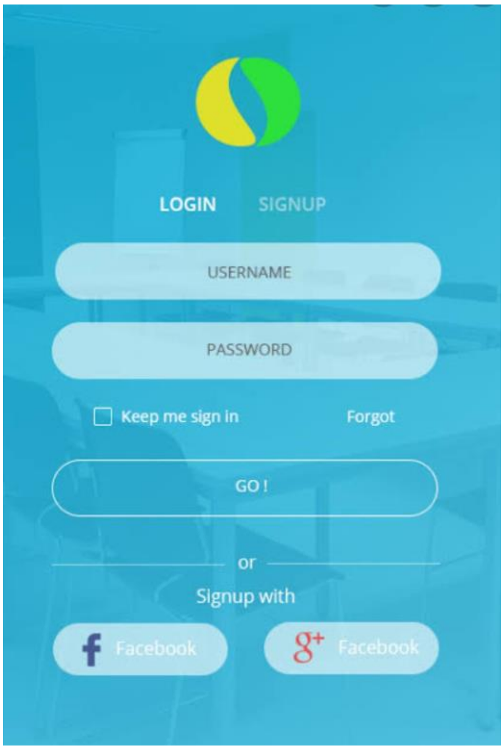
   Flutter is Google's UI toolkit for building beautiful, natively compiled applications for mobile, web, and desktop from a single codebase.

10. **PyTorch**

    An open source machine learning framework that accelerates the path from research prototyping to production deployment.

CHAPTER 5
PROJECT IMPLEMENTATION

5.1 User Interface



5.2 Google translator

```
from googletrans import Translator
translator = Translator()
translation=translator.translate('मेरी बिलीरुबिन अधिक है')

print(translation.origin, ' -> ', translation.text)
# <Translated src=ko dest=en text=Good evening. pronunciation=Good evening.>


```

```
from googletrans import Translator
translator = Translator()
translation=translator.translate('Where is this stomach pain located? ',dest='hindi')

print(translation.origin, ' -> ', translation.text)
```

```
Where is this stomach pain located?  -> जहां इस पेट दर्द स्थित है?
```

5.3 Output Of Model

Type in your question (512 word limit) and search search parameters

question_text: " my head is aching

search_similarity_by: answer

number_results_toReturn: 1

answer_only: False

```
1it [00:00,  5.27it/s]
Result  1
['i've been having constant headache for about 10 days now?']

Result  2
['migraine or sinusitis…']
```

5.4 Training

```python


def download_file_from_google_drive(id, destination):
    URL = "https://docs.google.com/uc?export=download"

    session = requests.Session()

    response = session.get(URL, params = { 'id' : id }, stream = True)
    token = get_confirm_token(response)

    if token:
        params = { 'id' : id, 'confirm' : token }
        response = session.get(URL, params = params, stream = True)

    save_response_content(response, destination)

def get_confirm_token(response):
    for key, value in response.cookies.items():
        if key.startswith('download_warning'):
            return value

    return None

def save_response_content(response, destination):
    CHUNK_SIZE = 32768

    with open(destination, "wb") as f:
        for chunk in response.iter_content(CHUNK_SIZE):
            if chunk: # filter out keep-alive new chunks
                f.write(chunk)

import os
import requests

import urllib.request

# Download the file from `url` and save it locally under `file_name`:
urllib.request.urlretrieve('https://github.com/naver/biobert-pretrained/releases/download/v1.0-pubmed-pmc/biobert_v1.0_pubmed_pmc.tar.gz', 'BioBert.tar.gz')

if not os.path.exists('BioBertFolder'):
    os.makedirs('BioBertFolder')
```

```

import os
import requests

import urllib.request

# Download the file from `url` and save it locally under `file_name`:
urllib.request.urlretrieve('https://github.com/naver/biobert-pretrained/releases/download/v1.0-pubmed-pmc/biobert_v1.0_pubmed_pmc.tar.gz', 'BioBert.tar.gz')

if not os.path.exists('BioBertFolder'):
    os.makedirs('BioBertFolder')

import tarfile
tar = tarfile.open("BioBert.tar.gz")
tar.extractall(path='BioBertFolder/')
tar.close()

file_id = '1uCXv6mQkFfpw5txGnVCsl93Db7t5Z2mp'

download_file_from_google_drive(file_id, 'Float16EmbeddingsExpanded5-27-19.pkl')

file_id = 'https://onedrive.live.com/download?cid=9DEDF3C1E2D7E77F&resid=9DEDF3C1E2D7E77F%2132792&authkey=AEQ8GtkcDbe3K98'

urllib.request.urlretrieve( file_id, 'DataAndCheckpoint.zip')

if not os.path.exists('newFolder'):
    os.makedirs('newFolder')

import zipfile
zip_ref = zipfile.ZipFile('DataAndCheckpoint.zip', 'r')
zip_ref.extractall('newFolder')
zip_ref.close()
```

5.5 Libraries Installation

```
#@title  { display-mode: "code" }

#To use CPU FAISS use
!wget  https://anaconda.org/pytorch/faiss-cpu/1.2.1/download/linux-64/faiss-cpu-1.2.1-py36_cuda9.0.176_1.tar.bz2
#To use GPU FAISS use
# !wget  https://anaconda.org/pytorch/faiss-gpu/1.2.1/download/linux-64/faiss-gpu-1.2.1-py36_cuda9.0.176_1.tar.bz2
!tar xvjf faiss-cpu-1.2.1-py36_cuda9.0.176_1.tar.bz2
!cp -r lib/python3.6/site-packages/* /usr/local/lib/python3.6/dist-packages/
!pip install mkl

!pip install tensorflow-gpu==2.0.0-alpha0
import tensorflow as tf
!pip install https://github.com/re-search/DocProduct/archive/v0.2.0_dev.zip
!pip install gpt2-estimator
!pip install pyarrow

```

CHAPTER 6
RESULT AND DISCUSSION

Our model converts successfully voice input into text then the native text into English language to feed it to the network. The BioBert network provides the remedy to the problem stated by the user of the application. The remedy is then translated into native language and then that text output is converted to voice output. The application was successfully deployed using Flutter to work flawlessly with Android and IOS. We have user login details served using Firebase. We understood how transformers work. We got a glimpse of Word2Vec a program to covert words to vector for use in Natural Language Processing. We learned how to use Google Translator libraries.We understood how flutter works and got to know its benefits.

CHAPTER 7
FUTURE SCOPE

Though our model was able to provide remedies for most of the queries provided to the model but it is not as good to be used in commercial health-care. It fails to give proper remedy for a fair share of queries. Moreover the remedy provided by the model cannot be trusted much. Health-care is very complex and we might need much more data to confidently map a particular symptom to a particular disease. With advancements in NLP these shortcomings can be compensated to a certain extent. With advancements in Computer Vision we might be able to take image data to recognize skin diseases and it can be helpful in other diseases also. Same technologies can be applied to feed the data of past prescriptions to understand symptoms better.

CHAPTER 8
CONCLUSION

A model with biomedical text predictions is successfully created to predict the cause of the problem a provide a suitable remedy. A mobile app based on flutter is proposed to take voice input of symptoms in native language like Hindi, Marathi, Telugu etc. The transformer model successfully outputs the remedy in native language.

CHAPTER 9

REFERENCES

Alsentzer,E. et al. (2019) Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA. pp. 72–78. Association for Computational Linguistics. https://www.acl web.org/anthology/W19-1909.

Bhasuran,B. and Natarajan,J. (2018) Automatic extraction of gene-disease associations from literature using joint ensemble learning. PLoS One, 13, e0200699.

Bravo,A` . et al. (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics, 16, 55.

Devlin,J. et al. (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA. pp. 4171–4186. Association for Computational Linguistics. https:// www.aclweb.org/anthology/N19-1423.

Dogan,R.I. et al. (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform., 47, 1–10. Gerner,M. et al. (2010) Linnaeus: a species name identification system for biomedical literature. BMC Bioinformatics, 11, 85.

Giorgi,J.M. and Bader,G.D. (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 34, 4087.